Multimed Tools Appl
DOI 10.1007/s11042-014-2056-5

Best practices for learning video concept detectors from social media examples

Svetlana Kordumova · Xirong Li · Cees G. M. Snoek

Received: 25 September 2013 / Revised: 3 April 2014 / Accepted: 25 April 2014

Abstract Learning video concept detectors from social media sources, such as Flickr images and YouTube videos, has the potential to address a wide variety of concept queries for video search. While the potential has been recognized by many, and progress on the topic has been impressive, we argue that key questions crucial to know *how to learn effective video concept detectors from social media examples?* remain open. As an initial attempt to answer these questions, we conduct an experimental study using a video search engine which is capable of learning concept detectors from social media examples, be it socially tagged videos or socially tagged images. Within the video search engine we investigate three strategies for positive example selection, three negative example selection strategies and three learning strategies. The performance is evaluated on the challenging TRECVID 2012 benchmark consisting of 600 h of Internet video. From the experiments we derive four best practices: (1) tagged images are a better source for learning video concepts than tagged videos, (2) selecting tag relevant positive training examples is always beneficial, (3) selecting relevant negative examples is advantageous and should be treated differently for video and image sources, and (4) learning concept detectors with selected relevant training data before learning is better then incorporating the relevance during the learning process. The best practices within our video search engine lead to state-of-the-art performance in the TRECVID 2013 benchmark for concept detection without manually provided annotations.

Keywords Video retrieval · Concept Detection · Social media

S. Kordumova · C. G. M. Snoek
Intelligent Systems Lab Amsterdam, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

S. Kordumova
e-mail: s.kordumova@uva.nl

C. G. M. Snoek
e-mail: cgmsnoek@uva.nl

X. Li (✉)
Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China,
No. 59 Zhongguancun Street, 100872 Beijing, China
e-mail: xirong@ruc.edu.cn

Published online: 23 May 2014

1 Introduction

Many videos are produced every day. Consider for example the amount of frames that are contained in the 100 h of Internet videos that are being uploaded to YouTube every minute. In order to pinpoint a specific query to arbitrary fragments of a video, semantic labels at a shot or even frame level are required. This leaves us no other choice but to devise machine tagging mechanisms that can detect visual concepts such as *animal*, *building*, and *snow* at the frame level [2, 32]. The state-of-the-art in video concept detection is to learn SVM classifiers from manually labeled frames represented by visual code features [9, 38]. However, the expense of manual labeling results in training examples with limited availability. As a consequence, the performance of concept detection is bounded to a narrow application domain where a limited array of concepts can be reliably detected [43].

The potential of automatically obtaining training data from the web for all possible video concepts one can think of is recognized by many [3, 17, 19, 27, 29, 32, 35, 41, 45]. To promote progress in this field of research, the well known benchmark for video retrieval TRECVID [32], even started a dedicated task for learning video concepts without manually provided annotations in 2012. To reach this goal, one promising line of research is to automatically acquire training examples from social media, where many socially tagged videos and images exist. Ulges et al. were among the first to learn video concepts from YouTube [35]. In their system, if social tags of a video match a given concept, all frames of the video are considered as positive examples of that concept. Setz and Snoek [29] conducted a pilot study on learning video concepts from Flickr images, directly treating images labeled with the concept as positives. However, social tags are known to be unreliable and often irrelevant with respect to the visual content they are describing [15, 18, 20, 36, 41, 45]. Hence, for learning meaningful concept detectors from social media, selecting appropriate examples from a proper data source is crucial.

As noted by Yang et al. [43], the performance of a concept detector could degenerate severely if the training and the test videos are from different genres, e.g., broadcast news and documentaries. However whether this will also hold true in a cross-source scenario, say applying image classifiers on video data, has not been investigated yet in the literature. We observe prevailing usage of socially tagged videos as the training source of choice [19, 34–36, 41, 44], even though selecting positive training examples from videos is more difficult than selecting positive examples from images. This is not only because video tags are noisy, but also the question of how to propagate tags to the frame level is still open [1, 19]. In fact, the question whether using tagged videos are more beneficial then simply using tagged images has not been addressed in the literature.

The vast collections of images and videos in social media offer many possibilities for learning visual detectors. Some researchers have focused on selecting relevant positive examples [18, 33, 36, 45], where others focus on selecting negative examples [15, 42]. The authors of [6, 19] incorporated the relevance in the learning methods. Among all these possibilities, one may wonder *How to learn effective video concept detectors from social media examples?*.

To acquire accurate positive examples from social media, a common approach is through a retrieval process, where socially tagged examples are first ranked in terms of their estimated relevance scores with respect to a given concept [18, 33, 36, 45]. Then, the top ranked proportion of the examples is preserved. In [18, 45], for instance, a fixed number of examples are selected for every concept, while [36] just tried a varying amount of frame fractions. Some even ignored the unreliability of the social labels, and directly used tagged videos [35] or tagged images [29] as positive examples. However, what strategy is the best for selecting positive examples to learn video concept detectors is unsettled.

For selecting negative training examples, a common strategy is to randomly sample images not tagged with the concept name [11, 19, 29, 35, 36, 45]. The classifiers usually misclassify negative examples which are visually similar to the positives. Therefore, inclusion of such misclassified and thus relevant negatives could lead to better concept detectors. Li et al. [17] select relevant negatives with an iterative bootstrap approach, where the top misclassified images are selected as relevant negatives. Yan et al. [42] select the bottom ranked examples based on some similarity metric to the positives as reliable negatives. Notice that both methods use expert-labeled positive training examples. To the best of our knowledge, learning video concept detectors with relevant negatives, relying fully on social tagged data, has not been investigated in the literature.

Instead of focusing on selecting relevant positive or negative examples [17, 19, 36, 42], some cope with the noisy social data by machine learning strategies that address the relevance during the detector training phase [5, 6, 19]. For example, the authors of [19] use multiple instance learning, where tagged videos are treated as bags, and the frames as instances. Hu et al. [5] learn an image ranking model from tagged image pairs with preference relations. Grauman et al. [6] learn relative tags on images with RankSVM [10], where ordered pairs are created from a ranked list of examples for the learning procedure. However, whether selecting relevant positive and negative examples a priori is better then relying on a learning strategy that optimizes the selection during training is unknown.

In this paper, we aim to identify best practices for learning video concept detectors from social media examples. We structure our paper as an experimental study with a complete system that learns visual concepts from socially labeled media examples, see Fig. 1. In an initial version of this work [13], we investigated positive example selection strategies only. In this paper we present a more comprehensive study by evaluating three negative example selection strategies and three learning strategies on a larger set of Internet videos. These new elements allow us to identify a set of best practices, and consequently lead to the winning system in the TRECVID 2013 benchmark for concept detection without annotations. Before describing our experimental study we first review related work.

2 Related work

We structure our literature review in terms of the choices one has to make for learning video concept detectors from social media examples. In Section 1 we briefly consider the data sources used. In Section 2 we review strategies for the selection of positive training examples.

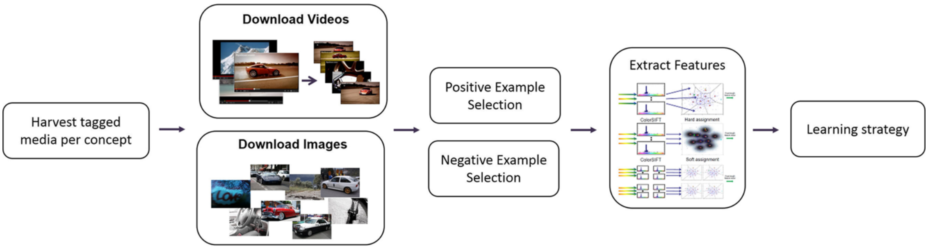


Fig. 1 In this paper we perform an experimental study with a video search engine to arrive at best practices for learning video concept detectors from social media examples

We consider the strategies for selecting negative training examples in Section 3. Learning strategies are summarized in Section 4.

2.1 Data sources

When harvesting visual training examples from social media, two sources have been used in the literature, namely tagged videos [19, 34–36, 44] and tagged images [3, 11, 17, 29, 45]. Ulges et al. [35] learn video concept detectors using YouTube videos by directly treating social tags as relevant labels for all video frames. Wang et al. [41] follow a similar approach, but require a set of manually labeled videos to bootstrap the learning process. Multiple instance learning was investigated in [19] treating socially tagged videos as bags, and frames as instances. Setz and Snoek [29] investigate whether Flickr images can be exploited as a direct training source for learning video concepts. Li et al. [15, 17] harvest relevant tagged images as positive and as negative training examples. Notice that none of the above works considers which *source* is a better choice. Therefore we raise the first research question:

> Research Question 1:
> What visual tagging source is most suited for selecting
> training examples for learning video concept detectors?

This study covers two representative approaches as baselines i.e., Ulges et al. [35] and Setz and Snoek [29], naming them as the ALL-FRAME and the ALL-IMAGE respectively.

2.2 Positive examples

For video it is easy to observe that even when tags are relevant at the video level, it does not imply that the tags are relevant for all shots and frames as well. Ulges et al. [36] model the relevance of video (key)frames with density estimation in the feature space. Since the density estimation is tag independent, for a video labeled with multiple tags, the method [36] cannot tell which tag is relevant w.r.t. which frame. Also, the prior information is unknown, so multiple fractions of the top ranked frames are used and the best fraction is selected on the expense of a manually labeled validation set. Since manual annotation is required for validating the prior values, this approach is unscalable for many concepts. Zhao et al. [44] suggest new tags on a video level by propagating tags from similar videos, ignoring the fact that not all frames of the video are relevant to the tag [19, 32, 36]. In order to select the relevant frames, we need to estimate tag relevance on a frame level. Given that a video is a composition of continuous frames, some have used the temporal dimension either as a feature to describe the video [26, 31], or to refine the end results of the positive examples with the scores of the temporal neighboring frames [19]. Since temporal features such as motion [40] and audio [8] are not applicable on images, we do not consider temporal information to ensure a fair comparison between the two sources at the image/frame level.

Positive example selection of tagged images has been addressed by many. For instance, the semantic field method [45] determines tag relevance in terms of tag-wise similarity. However if applied on videos, this approach may identify relevant tags for a video, but not for individual frames, since the semantic field works only on a video level. The authors of [3] detect spam, polysemous and synonym tags to filter out noisy tagged images. Their method relies on the assumption that users provide multiple tags per image. However the statistics presented in [30] show that in a large sampled Flickr collection 2/3 of the images have one to three tags at most. Thus relying on methods with tag co-occurrence may result in poor conclusions. Li et al. [15] learn the relevance of tags accompanying an image using a neighbor voting algorithm, by finding neighbors in the visual feature space. They exploit the observation that similar tags issued for similar images are reliable, by accumulating tag votes over visually similar images. Including a diverse set of visual features in the neighbor voting algorithm further improves the effectiveness of tag relevance [16]. An alternative to tag relevance is proposed by Liu et al. [20]. Their method is also founded on neighbor voting, but the neighbors are weighted with a Gaussian function. Notice that [20] restricts the neighbors to be images labeled with the tag, while [15] exploits the entire collection.

Since there are many possible choices one can make for selecting positive examples from tagged sources when learning video concept detectors, we raise the second question:

Research Question 2:
What strategy should be used for selecting positive examples
from tagged sources when learning video concept detectors?

In this paper we choose the neighbor voting algorithm [15] as the tag relevance measure because of its reported leading performance [37]. We evaluate three strategies for positive example selection, two based on tag relevance and one with random sampling.

2.3 Negative examples

While the problem of selecting positive examples from social media has been well investigated, the importance of negative examples has mostly been overlooked in the literature. Most concept detectors rely on simple random selection of images [11, 21, 29, 45] or videos [19, 35, 36] not tagged with the concept name.

The authors of [42] propose a *pseudo negative* selection of negative examples for video retrieval. The pseudo negative performs an iterative learning procedure by selecting reliable negatives from the worst matching examples identified by the classifier. Although sampling at the lower ranks probably yields reliable negatives, the classifier already ranked them correctly at the bottom, thus adding them to the training process can only yield minor improvements. Therefore, using other ranking measures for the examples, like tag relevance, and then selecting the bottom ranked examples is interesting to investigate.

Li et al. [17] propose a *negative bootstrap* approach, where relevant negatives are harvested from a large amount of social-tagged images. Since classifiers tend to misclassify negative examples which are visually similar to positive examples, including these relevant negatives boosts the performance. To that end, the negative bootstrap approach samples relevant negatives in an iterative manner. The misclassified negative examples are considered as relevant and thus leveraged in the consequent learning iteration. Using relevant negative sets result in visual classifiers with higher accuracy. While negative bootstrap is intended and shown to be effective for building image classifiers, its effectiveness for video concept detection is unknown. Moreover, the algorithm uses expert-labeled positive examples. To the best of our knowledge, a joint adventure of selecting relevant positive and relevant negative examples for video concept detection is non-existing. Therefore we raise the third research question:

Research Question 3:
What strategy should be used for selecting negative examples
from tagged sources when learning video concept detectors?

In this paper we evaluate three strategies for negative example selection, the negative bootstrap, the pseudo negative and random sampling from both tagged image and tagged videos.

2.4 Learning strategies

Since social tags can be ambiguous and subjective, tagged images and videos should be treated carefully when used as training data for learning video concept detectors. Many have focused on selecting relevant positive and negative examples before learning. Li et al. [19] bypass the selection problem by using Multiple Instance Learning (MIL) [28] to estimate the relevance of a video tag at the shot level. In a MIL setting, instances are organized into bags and it is the bags, instead of individual instances that are labeled for training. In [19] the tagged videos are considered as labeled bags, and the frames are the instances. If a video is labeled with a concept, all its frames are considered as positive, and if not, all its frames are considered as negative. However, the base MIL model [28] will hurt if videos are miss-labelled, and it is known that the social tags are imperfect and ambiguous. To overcome this obstacle, the authors of [19] calculate a tag correctness score and use it as a weight in the optimization function of the MI logistic regression. The tag correctness score is calculated by averaging the cosine similarity between the video tags and text snippets retrieved from a search engine with the target tag as a query. For a test frame, its predicted score is further smoothed using scores of its temporal neighbors. It is worth noting that for content based image retrieval, standard supervised learning is superior to MIL [25].

When learning a concept detector with multiple instance learning or Support Vector Machines [39], every positive and negative example used for training contributes equally in the optimization process. Dealing with noisy socially tagged data, it is potentially more robust to rely on a relative degree of relevance for each example. The initial idea of using relevant information of examples while learning by RankSVM, was originally proposed by Joachims [10] for web page ranking. Later the idea has been used together with multiple instance learning for finding regions of interest in the image [5]. The authors of [6, 12] use RankSVM to add relative information to tags, like *more then* or *less then*. Click data of image search engine is used with RankSVM to re-rank images [7].

Although different learning strategies have been investigated in many other tasks, we have not seen a comparative evaluation on what strategy is most suited for learning video concept detectors from social media examples. Therefore the forth question arises:

Research Question 4:
What learning strategy should be used for learning
video concept detectors from socially tagged data?

This study evaluates three learning strategies. We name the multiple instance learning approach of [19] as *Packed examples* learning, since frames are packed into bags. As *Paired examples* learning we refer to models learned from example pairs with preference relations [6]. We generate the ordered pairs based on the tag relevance scores of the examples. Moreover, we investigate whether bypassing the selection problem of relevant examples, as done in *Packed examples* and *Paired examples* learning is a better strategy than *Refined examples* learning, where relevant training examples are selected a priory and then used to train a one-versus-rest Support Vector Machines [39].

To answer the four research questions and obtain best practices, we structure our paper as an experimental study with a complete video search engine that learns concept detectors from social media examples. We identify four key components of such a system, 1) harvesting social media sources, 2) selection of positive training examples, 3) selection of negative training examples and 4) learning strategies to obtain concept detectors. The key components of our video search engine are highlighted in Fig. 1 and detailed next. Conceptual diagrams of the individual components are given in Figs. 2, 3 and 4, in sequence.

3 Experimental video search engine

3.1 Harvesting social media sources

We harvest two type of media sources, tagged videos and tagged images. The tagged videos are collected from YouTube, which is one of the most popular service for video sharing. The tagged images are selected from Flickr as one of the most popular sharing service for images. In this study, we consider 20 video concepts, covering objects like *Car, Plant* and scenes like *Outdoor, Landscape*. The set of 20 concepts was chosen to consist of single-word concepts only, for more precise tag relevance estimation, and restricted to those concepts with the most frequent number of positive examples in the data sets considered [15, 32].

Tagged videos In order to construct a diverse set, we collect videos retrieved by four distinct ranking criteria provided by YouTube, i.e., view count, relevance, date published and user rating. Consequently, we obtain for each of the 20 concepts four lists of retrieved videos and their metadata, containing videos id, tags, author, video duration etc. The date published ensures that new instances of concepts like *Car* and *Building* models are covered. We include in our dataset the top 50 retrieved videos from each of the four ordering criteria. Hence, for each concept we download the most viewed, most relevant, most recently uploaded and best rated videos. We shot segment each video and define the middle frame of each shot as its keyframe. Since we want to show the influence of frame selection, we maintain only those videos that have at least two shots. This process resulted in 200 h of web video and 130 K keyframes.

Tagged images We adopt the Flickr image collection from [15]. The images are of medium size with width or height fixed to 500 pixels. A subset is generated such that the number of images for each concept is maintained the same as the number of frames from the *Tagged Videos* dataset. In total, the *Tagged Images* collection has 130 K images.

Table 1 shows the number of videos and images labeled with a given concept in the two sources.

3.2 Positive example selection strategies

We investigate three selection strategies, see Fig. 2. Given a specific concept ω, we select positive training examples from the two sources described in Section 1. Let x be such an example in consideration. Depending on the source of training data, x is either an image labeled with ω or a frame extracted from a video labeled with ω. Let L_ω be the set of n examples labeled with ω.

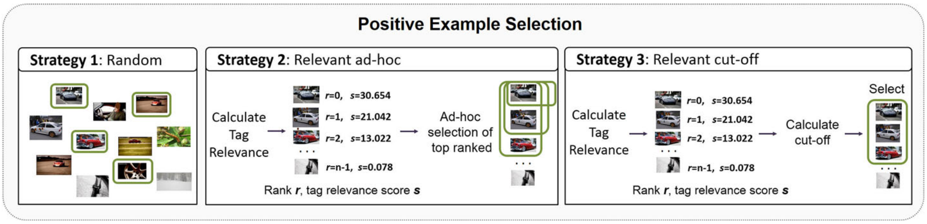


Fig. 2 Three strategies for positive training example selection for learning video concept detectors from social media: random, relevant ad-hoc and relevant cut-off

1. ***Random selection***. One straightforward selection strategy is to rely on the social tags and to randomly select k labeled examples with the concept ω. The set of randomly selected positive examples P_ω is

$$P_\omega = \left\{\forall x \in C \middle| C = \{x_1, x_2, \ldots, x_k\} \subset L_\omega \right\}. \tag{1}$$

2. ***Relevant ad-hoc selection***. As aforementioned, simply treating x as a positive example is problematic. Therefore another strategy is to calculate tag relevance per example, and then select the top ranked examples. For every example x labeled with concept ω, we use the multi-feature variant of the neighbor voting algorithm [16] to compute tag relevance scores. The tag relevance scores allow us to rank the examples such that the most relevant examples are deemed to be placed at the top. Therefore we can select the k top ranked examples as positives:

$$P_\omega = \left\{\forall x \in L_\omega \middle| rank(x) \in 0, 1, \ldots, k-1 \right\}. \tag{2}$$

3. ***Relevant cut-off selection***. Strategy 2. does not answer how many of the k examples should be selected from the ranked list. Validating the selection size comes with the expense of manual annotation for the validation set, making this approach unscalable for many concepts. Instead of setting an ad hoc threshold k [18, 45], or validating it with multiple runs [36] with the expense of manual annotation, in this work we consider a simple strategy to determine a cut-off automatically. To this end, we calculate which example x should be selected based on a Bayesian decision rule. For each example, we use s to denote its tag relevance score and r to denote the corresponding rank, where $r=0,\ldots,n-1$. We observe that the first ranked image after tag relevance is very often a positive response. Hence, we assume this first-hit example to be a correct reference point for estimating the probability of the other examples being positive. In that regard, we introduce a binary random variable y, where $y=1$ means x is positive, and 0 otherwise. The problem of positive example selection can then be simplified to estimating the conditional probability $p(y=1|x)$. With the Bayesian decision theorem, we define the selection rule simply as:

$$\begin{cases} \text{x is selected}, & \text{if } \dfrac{p\left(y=1\middle|x\right)}{p\left(y=0\middle|x\right)} > 1, \\ \text{unselected}, & \text{otherwise}. \end{cases} \tag{3}$$

For computing $p(y=1|x)$, we have access to two observations with respect to x, i.e., the relevance score s, and the corresponding rank r. Using a single observation is limited, since the scores are discriminative but less robust, while the quantized ranks tend to be more robust but less discriminative. Hence, we consider their combination, which should result in a better estimation of $p(y=1|x)$. Using probability algebra, we have $p(y|x)=p(y|s,r)=p(s,r|y)p(y)/p(s,r)$.

For $p(s,r|y)$, we make a practical simplification by estimating it through $p(s|y)\cdot p(r|y)$. We also expand $p(s|y)$ and $p(r|y)$ using Bayes' theorem. Accordingly, we rewrite

$$p\left(y\middle|x\right)\approx\frac{p\left(s\middle|y\right)p\left(r\middle|y\right)p(y)}{p(s,r)}=\frac{p\left(y\middle|s\right)p\left(y\middle|r\right)p(s)p(r)}{p(s,r)p(y)}, \tag{4}$$

For the unknown prior *p(y)*, we assume an uniform prior, and the decision function becomes

$$\frac{p\left(y=1\middle|x\right)}{p\left(y=0\middle|x\right)}=\frac{p\left(y=1\middle|s\right)p\left(y=1\middle|r\right)}{p\left(y=0\middle|s\right)p\left(y=0\middle|r\right)}. \tag{5}$$

With the intuition that examples with larger tag relevance scores and higher ranks are more likely to be positive, we approximate $p(y=1|s)$ as

$$p\left(y=1\middle|s\right)\approx\frac{s}{s_{max}}, \tag{6}$$

where s_{max} is the score of the top ranked example, and we compute $p(y=1|r)$ as

$$p\left(y=1\middle|r\right)\approx 1-\frac{r}{n}. \tag{7}$$

Accordingly, for concept ω we create a positive examples set P_ω as

$$P_\omega=\left\{\forall x\in L_\omega\middle|\frac{p\left(y=1\middle|x\right)}{p\left(y=0\middle|x\right)}>1\right\}. \tag{8}$$

In contrast to [18, 36, 45], the selection rule defined in Eq. (8) sets the cut-off automatically, without the need to tune k.

We visualize the three positive example selection strategies in Fig. 2.

3.3 Negative example selection strategies

We investigate three strategies for selecting negative examples for a specific concept ω, see Fig. 3. We notate the set of examples x not tagged with the concept ω as $\overline{L_\omega}$.

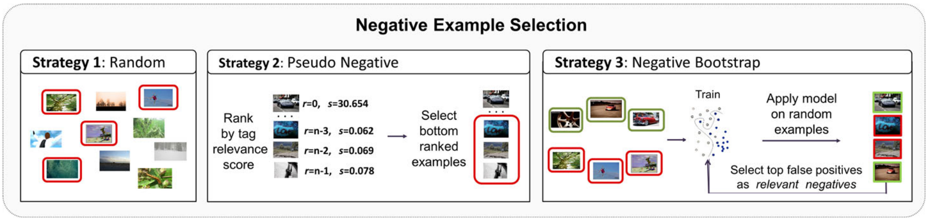

Negative Example Selection
Strategy 1: Random
Strategy 2: Pseudo Negative
Rank by tag relevance score
r=0, s=30.654
r=n-3, s=0.062
r=n-2, s=0.069
r=n-1, s=0.078
Select bottom ranked examples
Strategy 3: Negative Bootstrap
Train
Apply model on random examples
Select top false positives as relevant negatives

Fig. 3 Three strategies for negative training example selection for learning video concept detectors from social media: random, negative bootstrap and pseudo negative

1. ***Random selection***. The negative example selection strategy most commonly used in the literature is random sampling. We define the set of randomly selected q negative examples as

$$N_\omega = \left\{ \forall x \in C \middle| C = \{x_1, x_2, \ldots, x_q\} \subset \overline{L_\omega} \right\}. \tag{9}$$

2. ***Pseudo negative***. In order to acquire negatives which are more informative than the random negatives, we consider examples labeled with the concept name, but having low tag relevance scores. The basic idea is that images/videos tagged with the concept have low tag relevance because they differ from the other images tagged with the same concept, i.e. most of their visual neighbors do not have the same tag associated. However, the examples with low tag relevance score can still be informative since they are tagged with the concept name. Therefore it is interesting to investigate whether examples with low tag relevance if used as relevant negative examples will improve the performance of the concept detectors. The pseudo negative approach is inspired by to [42], in terms of selecting the bottom ranked examples, but with a difference in the ranking procedure. In [42] the bottom ranked images from a classification score are selected as negatives, in an iterative procedure. We select the negatives only in one step, using the tag relevance score to rank the examples tagged with the concept name, thus avoiding the expensive iterative learning. Using the decision rule calculated as in Eq. 5, we create the pseudo negative set as

$$N_\omega = \left\{ \forall x \in L_\omega \middle| \frac{p\left(y = 1 \middle| x\right)}{p\left(y = 0 \middle| x\right)} < 1 \right\}. \tag{10}$$

3. ***Negative bootstrap***. As mentioned in Section 3, using relevant negatives can improve the retrieval performance. Therefore as another selection strategy we evaluate the negative bootstrap [17] into our video search engine, for both tagged images and tagged videos. We keep the positive examples P fixed. The negatives are selected in an iterative learning procedure. In iteration t we denote the learned model as M_t, and a randomly sampled collection from $\overline{L_\omega}$ with γ examples as C_t. In each step t, the set of negative examples is

$$N_{\omega,t} = \left\{ \forall x \in \mathit{topmisclassified}(C_t, M_{t-1}, q) \middle| C_t = \{x_1, x_2, \ldots, x_\gamma\} \subset \overline{L_\omega} \right\}. \tag{11}$$

   Where $\mathit{topmisclassified}(C_t, M_{t-1}, q)$ are the top q elements from a random subset C_t with γ elements of $\overline{L_\omega}$ classified by the previous learned model M_{t-1} as positives. Moreover, in each step of the negative bootstrap process, relevant negatives are selected.

We visualize the three negative example selection strategies in Fig. 3.

3.4 Learning strategies

In order to answer the research question *what learning strategy should be used for learning video concept detectors from socially tagged data?* We consider three learning strategies, see Fig. 4. We investigate whether selection of relevant positive and negative examples before

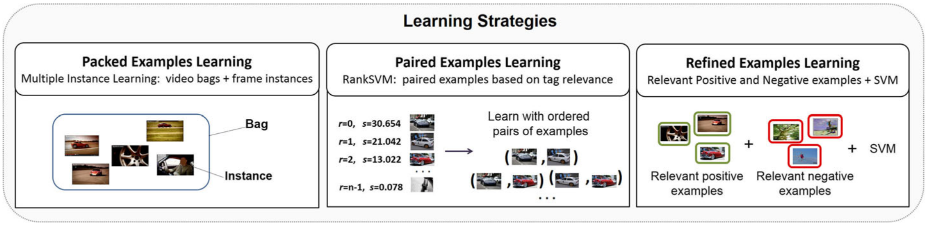

Learning Strategies
Packed Examples Learning
Multiple Instance Learning: video bags + frame instances
Bag
Instance
Paired Examples Learning
RankSVM: paired examples based on tag relevance
r=0, s=30.654
r=1, s=21.042
r=2, s=13.022
r=n-1, s=0.078
Learn with ordered pairs of examples
Refined Examples Learning
Relevant Positive and Negative examples + SVM
+
+ SVM
Relevant positive examples
Relevant negative examples

Fig. 4 Three learning strategies for video concept detectors with training data collected from socially tagged media: packed examples, paired examples and refined examples learning

learning concept detectors is more beneficial than using learning strategies that incorporate the relevance in the optimization function of the classifier.

1. ***Packed examples***. Multiple instance learning bypasses the selection problem of positive and negative examples when learning video concept detectors from tagged videos. The tagged videos are considered as labeled bags, and the frames are the instances. If a video is labeled with a concept, all its frames are considered as positive, and packed into a positive bag. If a video is not labeled with a concept, it is considered as a negative bag, and its frames as negative instances. We employ the implementation of [28] into our video search engine, with the additional weight for video tag relevance as in [19].

Table 1 Number of positive examples for 20 concepts in both the Tagged Videos and Tagged Images datasets

Concept	YouTube videos	Video frames	Flickr images
Animal	191	7,506	7,506
Beach	185	6,554	6,554
Building	197	5,207	5,207
Car	192	8,508	8,508
Child	174	6,938	6,938
City	169	8,326	8,326
Face	169	7,210	7,210
Hand	179	5,835	5,835
Landscape	181	4,005	4,005
Mountain	181	6,445	6,445
Oceans	150	5,755	5,755
Outdoor	184	6,103	6,103
Plant	151	5,139	5,139
Road	170	8,047	8,047
Sky	154	6,261	6,261
Snow	177	7,467	7,467
Sports	196	6,685	6,685
Streets	141	7,493	7,493
Trees	158	5,805	5,805
Vehicle	184	4,621	4,621

2. ***Paired examples***. RankSVM is a pairwise learning to rank approach that learns the ranking function from ordered pairs of examples. To train concept detectors for each concept, we derive a set of ordered visual examples:

$$R_\omega = \left\{ \forall \left(x_i, x_j\right) \in \left(L_\omega, L_\omega\right) \middle| tagrelevance(x_i) > tagrelevance\left(x_j\right) \right\}. \qquad (12)$$

   Each pair (x_i,x_j) denotes that x_i is more relevant than x_j since it has a higher tag relevance score. We use the implementation of [10] to learn video concept detectors both from tagged images and tagged videos.
3. ***Refined examples***. We select relevant positive and relevant negative examples to learn a video concept detector. As classifier we employ a one-vs-all Support Vector Machines [39] with the fast histogram intersection kernel [23] for its high efficiency. The SVM models are optimized by 3-fold cross validation.

We visualize the three learning strategies in Fig. 4.

4 Experiments

In order to arrive at best practices for the four key components described in Section 3, we conduct the following four experiments.

4.1 Experiment 1: what source?

In this experiment we address the research question *what visual tagging source is most suited for selecting positive training examples to learn video concepts?*. We use the *Tagged Videos* and the *Tagged Images* datasets described in Section 1 as instantiations of two distinct visual tagging sources. We learn concept detectors using the ALL-FRAME [21] and the ALL-IMAGE [14] scenarios separately, and compare their performance.

4.2 Experiment 2: what positive examples?

Besides the ALL-FRAME and the ALL-IMAGE baselines, we evaluate the three positive example selection strategies described in Section 2 to answer the question *what strategy should be used for selecting positive examples from tagged sources when learning video concept detectors*?. We employ the strategies for both the *Tagged Videos* and the *Tagged Images* datasets.

1. ***Random***. We randomly sample positive frames from the socially tagged sources for each concept. The set of positive examples is formalized in Eq. (1). We vary the number k of sampled positive frames to investigate their influence.
2. ***Relevant ad-hoc***. We rank the candidate positive examples by their tag relevance scores in descending order. For each concept, an ad-hoc fractions of the top ranked examples are selected as positive training examples. The set of positive frames is formalized in Eq. (2). We vary the fraction k=0.1, 0.2,…,0.9 to investigate the importance of making a proper cut-off.
3. ***Relevant cut-off***. We rely on the same ranking of the examples as in strategy 2, but here we automate the selection with the simple Bayes approach given in Section 2. The set of positive frames is defined in Eq. (8). Due to the large variance of the tag relevance scores for tagged images, we smooth the scores using the common logarithm.

Since we evaluate the positive selection strategies in this experiment, we choose the most simple and commonly used strategy for negatives, and keep it fixed. The same set of random sampled negatives, described in Section 3, were used for all strategies and the All-FRAME and ALL-IMAGE baselines, sampled form the *Tagged Images* and *Tagged Videos* sources respectively. Following the common procedure in the literature [17, 42] we set the size of the negative set to be the same as the number of positive examples. We refer to Table 1 for the number of positive examples per concept.

4.3 Experiment 3: what negative examples?

To answer *what strategy is most suited for selecting negative examples from social media?* we investigate the following three strategies for negative examples selection, detailed in Section 3. We employ the strategies on both the *Tagged Videos* and the *Tagged Images* datasets.

1. ***Random***. We randomly sample negative frames from the *Tagged Videos* for each concept. We vary the number of sampled negative frames to investigate their influence. The set of negative frames is formalized in Eq. (9).
2. ***Pseudo Negative***. We calculate a tag relevance score for each keyframe before ranking. A fraction of the *bottom* ranked frames per concept are selected as the negative training examples. The set of negative frames is formalized in Eq. (10).
3. ***Negative Bootstrap***. We employ an iterative procedure for selecting relevant negatives [17]. The set of negative frames is formalized in Eq. (11).

The same set of positives was used for all negative strategies, sampled from the corresponding tagged sources. The automatic cut-off strategy described in Section 2 was used. Thus, the relevant positive examples are selected once in a probabilistic manner, instead of setting an add-hoc threshold.

4.4 Experiment 4: what learning strategy?

We compare the performance of three learning strategies described in Section 4 to answer the question *what learning strategy should be used for learning video concept detectors from socially tagged data?*. We investigate whether *Refined examples* learning of selected relevant positive and negative training examples is better than methods that avoid prior selection of relevant examples. In *Packed examples* learning videos tagged with the concept name are considered as positive bags, and all its frames as positive instances. In *Paired examples* learning, like RankSVM, ordered pairs of examples are generated prior to learning. In both these methods the selection of relevant training data is ignored. In this experiment we compare the performance of *Paired examples* and *Packed examples* with *Refined examples* learning.

4.5 Test set

As test data we adopt the challenging internet video collection from the TRECVID 2012 benchmark [32]. We use the development dataset provided for the Semantic Indexing task, which consists of 600 h of Internet Archive videos, having 400,682 shots. Each shot is represented with a single keyframe. The dataset comes with ground truth annotations on a keyframe level, including the 20 concepts identified in our experiments. We evaluate and report all experiments on this test set.

4.6 Implementation details

Concept detection features. We employ a standard bag of visual codes pipeline to train video concept detectors [38]. We compute SIFT descriptors [22] at dense sampled points, at every 6 pixels for two scales. As visual features we employ a spatial pyramid of 1×1 and 1×3. The codebook size is 4,096, constructed with *k*-means clustering.

Tag relevance features. We follow the implementation from [16] and use three visual features: Color64, CSLBP and GIST. Color64 is a 64-d global feature, combining the 44-d color correlogram, the 14-d texture moments, and the 6-d RGB color moments [14]. CSLBP is a 80-d center-symmetric local binary pattern histogram [4], capturing local texture distributions. GIST is a 960-d feature describing dominant spatial structures of a scene by a set of perceptual measures such as naturalness, openness, and roughness [24].

Evaluation criteria. We adopt average precision to report individual concept detection accuracy, a common approach in the video retrieval literature [32]. We report mean average precision to evaluate the overall detection performance for all concepts.

Table 2 Experiment 1: what source?

Concept	Tagged videos	Tagged images
Animal	0.078	**0.122**
Beach	0.158	**0.359**
Building	0.334	**0.500**
Car	0.157	**0.230**
Child	0.069	**0.118**
City	0.064	**0.131**
Face	0.496	**0.606**
Hand	0.114	**0.175**
Landscape	0.207	**0.567**
Mountain	0.048	**0.516**
Oceans	0.129	**0.481**
Outdoor	**0.816**	0.722
Plant	0.188	**0.270**
Road	0.186	**0.427**
Sky	0.258	**0.621**
Snow	0.063	**0.273**
Sports	0.129	**0.149**
Streets	0.099	**0.183**
Trees	0.470	**0.693**
Vehicle	0.221	**0.351**
mAP	0.214	**0.375**

As first research question we investigate what visual tagging source, tagged images or tagged videos, is most suited for selecting training examples for learning video concept detectors. It leads to our first best practice: To learn video concept detectors from socially tagged media, tagged images should be used

Data in bold indicates the top performers for individual concepts

5 Results

5.1 Experiment 1: what source?

We summarize the results of Experiment 1 in Table 2. We observe that better concept detectors are obtained from tagged images than from the tagged videos. When simply using all available tagged images for training, we obtain an mAP of 0.375, while the video alternative obtains an mAP of 0.214. In terms of an absolute difference tagged videos have a significant 16.1 % lower mAP over tagged images. We explain the difference by the observation that the user tagging accuracy of images is more accurate when compared to videos. Moreover, when a tag is assigned to a video, often only a small fraction of the video content is relevant with respect to the tag. Consequently, given the same number of positive training examples, the positive set selected with our relevant selection from the image source contains more genuine positives than the positive set from the video source, as exemplified in Fig. 2. As our first best practice, we conclude that tagged images are a better source than tagged videos for learning video concept detectors.

5.2 Experiment 2: what positive examples?

Figure 5 shows the performance curve of concept detectors derived after positive example selection by the three strategies, for tagged videos and tagged images. In line with the results of Experiment 1, using images as training source is better than video for all strategies. Relevance selection is a better strategy than randomly selecting frames or images for all cut-offs in both sources. For the best possible relevant image selection (fraction=0.6), we gain a 5.4 % relative improvement over the same amount of randomly selected images. For the video alternative the difference is even more convincing. In that case we obtain the best possible result for a fraction of 0.1, with a 58.9 % relative improvement over random selection. There is a clear decrease in performance as we select larger fractions of frames, see *2. Relevant ad-hoc Frames* in Fig. 5. Since larger fractions contain more noisy data, i.e. frames that also have low tag relevance score, the classifier learns less accurate models. In the case of tagged images there is less noise compared to tagged videos, since the tag is

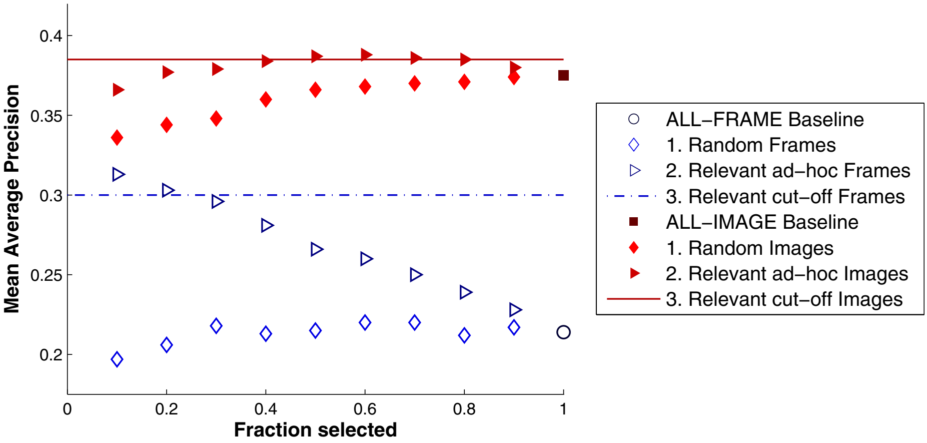


Fig. 5 Experiment 2: What positive examples? Training concept detectors after tag relevance selection, with either video frames or images is always beneficial. The simple cut-off selection is a good approximation of the best possible selected fraction for both video frames and images

Table 3 Experiment 2: what positive examples?

Concept	Strategies using tagged videos				Strategies using tagged images			
	1. Random	2. Add-hoc least	2. Add-hoc best	3. Cut-off	1. Random	2. Add-hoc least	2. Add-hoc best	3. Cut-off
Animal	0.060	0.083	0.106	0.102	0.116	0.122	0.140	**0.140**
Beach	0.208	0.146	0.256	0.245	0.324	0.306	0.340	**0.324**
Building	0.254	0.347	0.406	0.401	0.486	0.495	**0.544**	0.541
Car	0.128	0.168	0.249	0.237	0.234	**0.322**	0.293	0.308
Child	0.062	0.075	0.107	0.093	0.117	**0.140**	0.130	0.131
City	0.046	0.071	0.153	0.106	0.147	**0.277**	0.195	0.196
Face	0.400	0.503	0.562	0.536	0.594	0.594	**0.619**	0.602
Hand	0.125	0.112	0.116	0.117	**0.177**	0.152	0.163	0.159
Landscape	0.208	0.222	0.410	0.341	0.561	0.554	**0.580**	0.574
Mountain	0.043	0.057	0.255	0.214	0.490	0.468	**0.515**	0.506
Oceans	0.100	0.220	0.410	0.398	0.479	0.474	**0.516**	0.483
Outdoor	0.723	0.821	0.825	**0.831**	0.719	0.769	0.753	0.754
Plant	0.171	0.184	**0.286**	0.283	0.254	0.230	0.259	0.255
Road	0.242	0.235	0.327	0.358	0.408	0.372	**0.418**	0.416
Sky	0.258	0.277	0.371	0.391	0.623	0.412	0.601	**0.629**
Snow	0.050	0.077	0.138	0.127	**0.262**	0.198	0.259	0.242
Sports	0.122	0.137	0.204	0.184	0.145	**0.258**	0.215	0.225
Streets	0.111	0.102	0.169	0.124	**0.192**	0.191	0.184	0.184
Trees	0.458	0.490	0.580	0.578	**0.688**	0.609	0.660	0.648
Vehicle	0.183	0.227	0.339	0.329	0.348	0.381	0.381	**0.386**
mAP	0.197	0.228	0.313	0.300	0.368	0.366	**0.388**	0.385

With the second research question we investigate what strategy should be used for selecting positive examples from social media. We show results per concepts for three strategies, 1. Random, 2. Relevant ad-hoc selection with the best and least performance, and 3 Relevant cut-off selection. It leads to our second best practice: for learning video concept detectors, as positive examples use relevant cut-off selection of tagged images

Data in bold indicates the top performers for individual concepts

directly appointed to the visual content of the image. Consequently, for images larger fractions show better performance, see *2. Relevant ad-hoc Images* in Fig. 5. In this case, selection of smaller fractions also ignores relevant images, which results in less accurate classifiers. Apart from the gain in performance, it is also important to note that by selecting relevant subsets from the collection we reduce the number of training examples, which speeds up concept detector training.

The simple cut-off approach for automatic selection of positive examples, approximates the best selection quite closely (see the solid and dashed lines in Fig. 5). In case of image selection it outperforms the best possible ad-hoc relevant image selection for 9 concepts even, and the worst performing ad-hoc selection in 14 concepts (see Table 3). We conclude that selecting relevant positive images and video frames is needed when learning concepts from the web. The selection cut off can be estimated automatically. We employed a simple Bayes approach, although more extensive sampling methods can be incorporated in the future. For now, as second best practice we suggest to rely on the relevant cut-off selection strategy described in Section 2 for selecting relevant examples from socially tagged media.

Figure 6 shows some relevant positive examples automatically selected from social media.

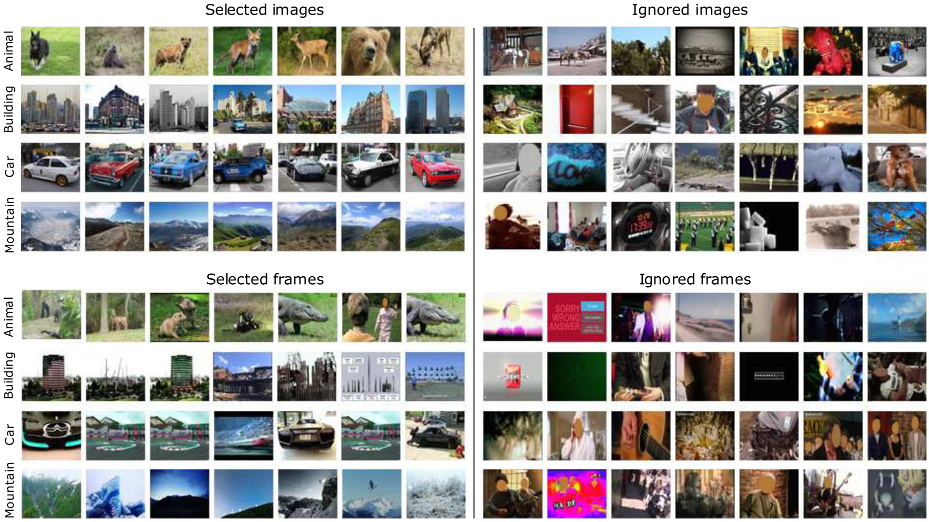


Fig. 6 Relevant positive training examples selected from social media. *Left columns* show the selected images and frames, while the *right columns* show images and videos labeled with a given concept but discarded by the relevant positive selection strategy

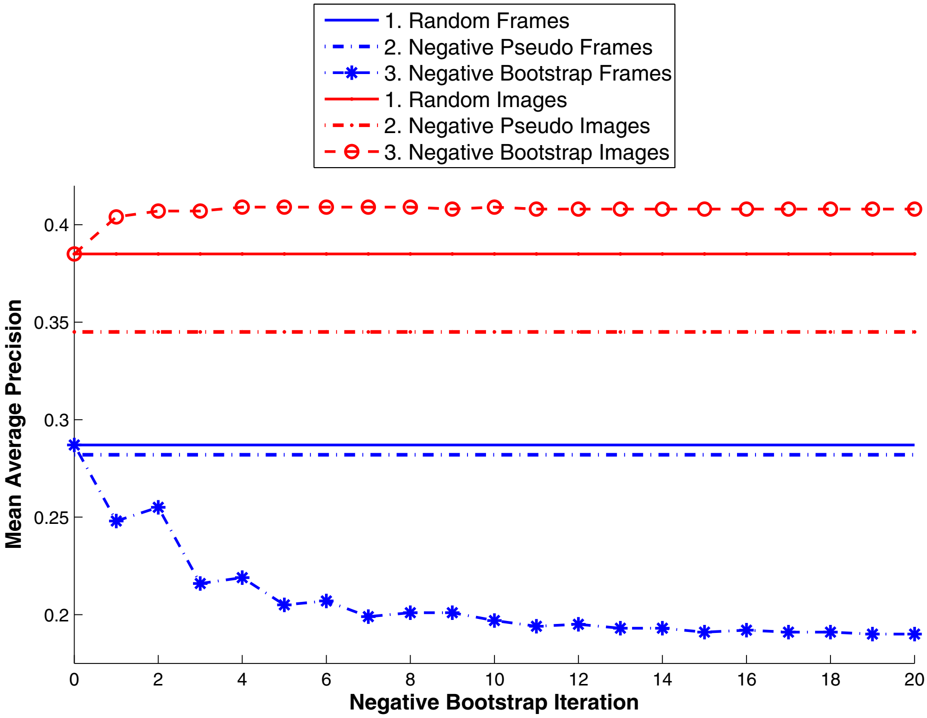


Fig. 7 Experiment 3: What Negative Examples? Selecting relevant negative examples from tagged images with the negative bootstrap improves over random sampling. The result converges after only five iterations. For tagged videos random sampling is the best choice because it suffers less from selecting false negatives than the pseudo negative and the negative bootstrap

5.3 Experiment 3: what negative examples?

In Fig. 7 we show the performance for the three negative example selection strategies for both tagged sources, images and videos. When using tagged images, our results confirm the finding of [17]. Compared to random sampling the negative bootstrap wins in 14 out of 20 concepts, with a 6 % relative improvement in mAP over random sampling. The pseudo negative strategy shows lower performance compared to random sampling, with 12.6 % relative decrease. By examining the selected negatives we find that the pseudo negative tends to select some positive examples as negatives, as shown in Fig. 8. Obviously, such false negative examples hurt the learning process.

In the case of tagged videos the situation is different. Random sampling is better then both negative bootstrap and pseudo negative. Since videos can have very diverse visual content, even if some video is not tagged with a concept name, it may contain some frames relevant to the concept. The negative bootstrap and the pseudo negative tend to select positive frames as negatives, see Fig. 8, which in terms of mAP result in less accurate performance over random sampling. However, the negative bootstrap wins over random sampling for 4 concepts, and the pseudo negative for 9 out of 20 concepts, as shown in Table 4. Therefore, selecting the most appropriate strategy for the concepts separately is interesting to be investigated for tagged videos.

Since selecting relevant negative examples from tagged images shows much better performance, our third best practice is to use negative bootstrap of tagged images to select negative training examples for learning video concept detectors.

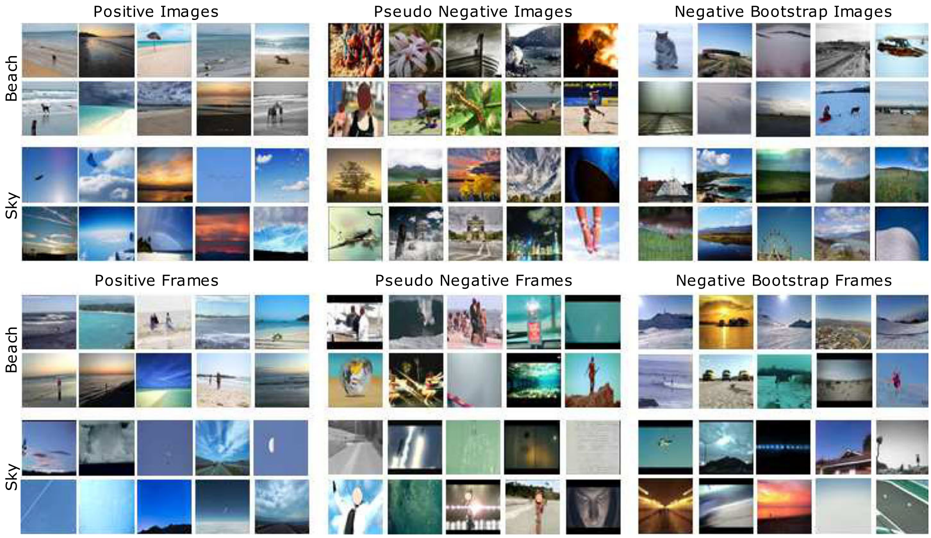


Fig. 8 Training examples selected from tagged images and tagged videos. *Left column* shows selected positive examples, the *middle column* shows selected examples with the pseudo negative, and the *right column* shows negative examples selected with the negative bootstrap. For the concept beach with the negative bootstrap we see selected images of snow or a landscape, i.e. images which are most likely to be mistaken by the concept detector. Relevant frames are also selected from videos, but also some false negatives. The pseudo negative method tends to select false negative examples of beach for both images and videos. For the sky concept, since many outdoor images or videos contain sky, but are not explicitly tagged with sky, both pseudo negative and negative bootstrap have a tendency to select false negatives

5.4 Experiment 4: what learning strategy?

In Table 5 we compare the three learning strategies described in Section 4 together with the ALL-FRAME [35] and ALL-IMAGE [29] baselines. When we follow the ALL-FRAME baseline [35], which simply uses all frames of tagged videos, rather than selecting relevant positive or negative frames, we obtain better results than the packed frames learning [19], with 17.6 % relative improvement in terms of mAP. The refined examples strategy for tagged videos also shows lower performance than the baseline, mostly because of the selection of false negative examples, as explained in more detail in Section 3. The paired examples learning has best performance for tagged videos, showing that using relative information of frames by creating ordered pairs is better than preselecting relevant positive and negative examples. When relying on images as training source for concept detectors we observe different behavior. Learning with refined examples reaches an mAP of 0.408, clearly improving over all other learning strategies, and with best result for 13 out of 20 concepts, see Table 5. We conclude with the best practice that training concept detectors from a

Table 4 Experiment 3: what negative examples?

Concept	Strategies using tagged videos			Strategies using tagged images		
	1. Random	2. Pseudo negative	3. Negative bootstrap	1. Random	2. Pseudo negative	3. Negative bootstrap
Animal	0.101	0.095	0.070	0.140	0.099	**0.157**
Beach	0.225	0.226	0.103	0.324	0.260	**0.361**
Building	0.383	0.308	0.294	0.541	0.447	**0.575**
Car	0.221	0.191	0.294	0.308	0.291	**0.384**
Child	0.096	0.111	0.068	**0.131**	0.121	0.126
City	0.100	0.131	0.128	0.196	0.216	**0.238**
Face	0.548	0.528	0.422	0.602	0.531	**0.672**
Hand	0.120	0.110	0.133	**0.159**	0.106	0.141
Landscape	0.314	0.331	0.122	0.574	0.537	**0.611**
Mountain	0.196	0.220	0.037	0.506	0.388	**0.539**
Oceans	0.358	0.390	0.037	0.483	0.448	**0.541**
Outdoor	0.815	0.801	0.562	0.754	**0.817**	0.769
Plant	0.274	**0.309**	0.161	0.255	0.238	0.265
Road	0.336	0.302	0.190	0.416	0.301	**0.474**
Sky	0.351	0.368	0.176	**0.629**	0.587	0.619
Snow	0.109	0.105	0.087	**0.242**	0.139	0.229
Sports	0.170	0.146	0.153	0.225	**0.232**	0.171
Streets	0.113	0.125	0.129	0.184	0.186	**0.214**
Trees	0.576	0.570	0.436	**0.648**	0.582	0.646
Vehicle	0.339	0.282	0.202	0.386	0.380	**0.431**
mAP	0.287	0.282	0.190	0.385	0.345	**0.408**

As third research question we investigate what strategy should be used for selecting negative examples from tagged sources when learning video concept detectors. It leads to our third best practice: bootstrapping relevant negatives from tagged images

Data in bold indicates the top performers for individual concepts

tagged image source using refined examples, i.e. relevant positive and negative images, is the best learning strategy.

5.5 Benchmark comparison

By taking all best practices into account, we participated in the NIST TRECVID 2013 [32] benchmark, using our video search engine. We considered the no annotation version of the task, and summarize results for all evaluated concepts in Fig. 9. Our detectors obtain the best performance for 28 out of 38 concepts.

6 Conclusion

This paper strives to answer *how to learn effective video concept detectors from social media examples* through a systematic empirical study. Supported by experiments on a present day

Table 5 Experiment 4: what learning strategies?

Concept	Strategies using tagged videos				Strategies using tagged images		
	ALL-FRAME	Packed examples	Paired examples	Refined examples	ALL-IMAGE	Paired examples	Refined examples
Animal	0.087	0.061	0.078	0.070	0.122	0.108	**0.157**
Beach	0.243	0.055	0.158	0.103	0.359	0.326	**0.361**
Building	0.323	0.154	0.334	0.294	0.500	0.496	**0.575**
Car	0.165	0.080	0.157	0.294	0.230	0.218	**0.384**
Child	0.083	0.080	0.069	0.068	0.118	0.111	**0.126**
City	0.091	0.048	0.064	0.128	0.131	0.160	**0.238**
Face	0.491	0.313	0.496	0.422	0.606	0.564	**0.672**
Hand	0.118	0.104	0.114	0.133	**0.175**	0.152	0.141
Landscape	0.276	0.099	0.207	0.122	0.567	0.520	**0.611**
Mountain	0.104	0.033	0.048	0.037	0.516	0.478	**0.539**
Oceans	0.288	0.092	0.129	0.037	0.481	0.463	**0.541**
Outdoor	0.815	0.743	**0.816**	0.562	0.722	0.750	0.769
Plant	0.244	0.214	0.188	0.161	**0.270**	0.249	0.265
Road	0.282	0.215	0.186	0.190	0.427	0.379	**0.474**
Sky	0.349	0.206	0.258	0.176	0.621	**0.626**	0.619
Snow	0.102	0.038	0.063	0.087	**0.273**	0.219	0.229
Sports	0.150	0.109	0.129	0.153	0.149	**0.187**	0.171
Streets	0.127	0.137	0.099	0.129	0.183	0.180	**0.214**
Trees	0.544	0.459	0.470	0.436	**0.693**	0.646	0.646
Vehicle	0.296	0.411	0.221	0.202	0.351	0.311	**0.431**
mAP	0.214	0.182	0.259	0.190	0.375	0.357	**0.408**

As forth research question we investigate what learning strategy is most suited for learning video concept detectors from social media examples. We show results for the three learning strategies and baselines for tagged images and tagged videos. We recommend a forth best practice: To use SVM with refined examples learning from tagged images, by selecting relevant negative and positive training examples

Data in bold indicates the top performers for individual concepts

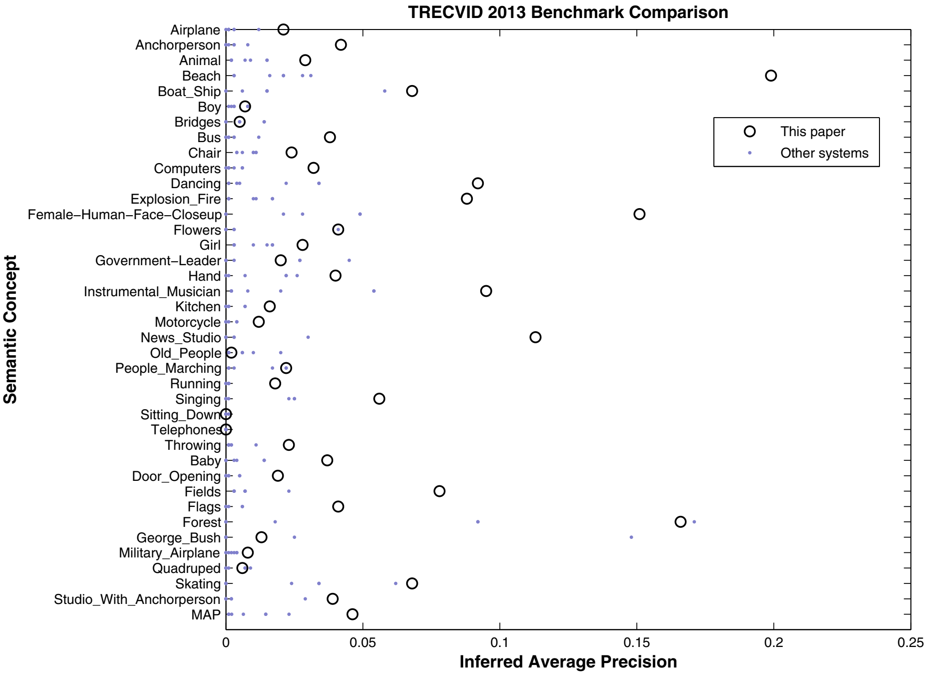


Fig. 9 Results per concept from the TRECVID 2013 Semantic Indexing with No Annotations task. The recommended best practices of this paper within our video search engine achieved best result for 28 out of 38 concepts, compared to the other submitted runs in the same task

testbed, our major findings are: 1) Tagged images are the preferred choice as training source, when compared to tagged videos. Concept detectors trained on tagged images surpass their counterparts trained on tagged videos, with an absolute improvement of approximately 16 % in terms of mAP. While images may not be one's first option due to source change, we find that the better annotation quality let them beat videos with ease. 2) For both tagged videos and tagged images, selecting positive examples from those with the largest tag relevance scores is superior to getting positives at random. We show that the cut-off of the top ranked relevant examples can be simply estimated, obtaining a near-optimal result. 3) The selection of negative training examples is also important. With negative bootstrap of tagged images we gain 6 % relative improvement over simple random selection. For videos random sampling is the best choice. 4) We show that preselecting relevant positive and negative training examples beforehand and training support vector machine is more beneficial then strategies where the relevance of the tagged training data is considered during learning. Below we summarize our best practices:

Best Practices:

1. Tagged images are a better source then tagged videos for learning video concept detectors.
2. Positive examples with relevant cut-off of tagged images show best performance.
3. Relevant negatives are best selected with negative bootstrap of tagged images.
4. For learning we recommend SVM with relevant positive and negative examples.

When the best practices of this paper are combined into a single video search engine it results in state of the art video concept detection without the need of any predefined manual annotations.

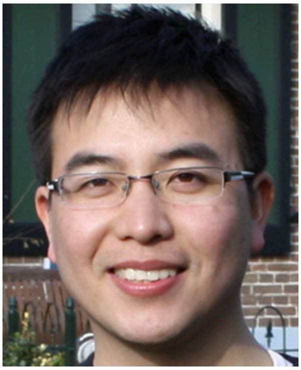
Acknowledgments This research is supported by the STW STORY project, the Dutch national program COMMIT, the Chinese NSFC (No. 61303184), SRFDP (No. 20130004120006), the Fundamental Research Funds for the Central Universities, the Research Funds of Renmin University of China (No. 14XNLQ01), and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

1. Ballan L, Bertini M, Del Bimbo A, Serra G (2011) Enriching and localizing semantic tags in internet videos. In: MM 1541–1544
2. Chang S-F, Ellis D, Jiang W, Lee K, Yanagawa A, Loui AC, Luo J (2007) Large-scale multimodal semantic concept detection for consumer video. In: MIR 255–264
3. Fan J, Shen Y, Zhou N, Gao Y (2010) Harvesting large-scale weakly-tagged image databases from the web. In: CVPR 802–809
4. Heikkila M, Pietikainen M, Schmid C (2009) Description of interest regions with local binary patterns. In: PR 42(3):425–436
5. Hu Y, Li M, Yu N (2008) Multiple-instance ranking: learning to rank images for image retrieval. In: CVPR 1–8
6. Hwang SJ, Grauman K (2012) Learning the relative importance of objects from tagged images for retrieval and cross-modal search. In: IJCV 100(2):134–153
7. Jain V, Varma M (2011) Learning to re-rank: query-dependent image re-ranking using click data. In: WWW 277–286
8. Jiang W, Cotton CV, Chang S-F, Ellis D, Loui AC (2009) Short-term audio-visual atoms for generic video concept classification. In: MM. doi:10.1145/1631272.1631277
9. Jiang Y-G, Yang J, Ngo C-W, Hauptmann A (2010) Representations of keypoint-based semantic concept detection: a comprehensive study. In: TMM 12(1):42–53
10. Joachims T (2002) Optimizing search engines using clickthrough data. In: SIGKDD 133–142
11. Kennedy LS, Chang S-F, Kozintsev IV (2006) To search or to label?: predicting the performance of search-based automatic image classifiers. In: MIR 249–258
12. Kim J, Pavlovic V (2012) Attribute rating for classification of visual objects. In: ICPR 1611–1614
13. Kordumova S, Li X, Snoek CGM (2013) Evaluating sources and strategies for learning video concepts from social media. In: CBMI 91–96
14. Li M (2007) Texture moment for content-based image retrieval. In: ICME 508–511
15. Li X, Snoek CGM, Worring M (2009) Learning social tag relevance by neighbor voting. In: TMM 11(7):1310–1322
16. Li X, Snoek CGM, Worring M (2010) Unsupervised multi-feature tag relevance learning for social image retrieval. In: CIVR 10–17
17. Li X, Snoek CGM, Worring M, Koelma DC, Smeulders AWM (2013) Bootstrapping visual categorization with relevant negatives. In: TMM 15(4):933–945
18. Li X, Snoek CGM, Worring M, Smeulders AWM (2012) Harvesting social images for bi-concept search. In: TMM 14(4):1091–1104
19. Li G, Wang M, Zheng Y-T, Li H, Zha Z-J, Chua T-S (2011) Shottagger: tag location for internet videos. In: ICMR. doi:10.1145/1991996.1992033
20. Liu D, Hua X, Yang L, Wang M, Zhang H (2009) Tag ranking. In: WWW 351–360
21. Liu Y, Xu D, Tsang IW-H, Luo J (2011) Textual query of personal photos facilitated by large-scale web data. In: PAMI 33(5):1022–1036
22. Lowe DG (2003) Distinctive image features from scale-invariant keypoints. In: IJCV 60(2):91–110
23. Maji S, Berg A, Malik J (2008) Classification using intersection kernel support vector machines is efficient. In: CVPR 1–8
24. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. In: IJCV 42(3):145–175

25. Ray S, Craven M (2005) Supervised versus multiple instance learning: an empirical comparison. In ICML 697–704
26. Schindler G, Zitnick L, Brown M (2008) Internet video category recognition. CVPR. doi:10.1109/CVPRW.2008.4562960
27. Schroff F, Criminisi A, Zisserman A (2007) Harvesting image databases from the web. In: ICCV 33(4):754–66
28. Settles B, Craven M, Ray S (2008) Multiple-instance active learning. In: NIPS 1289–1296
29. Setz A, Snoek CGM (2009) Can social tagged images aid concept-based video search? In: ICME 1460–1463
30. Sigurbjörnsson B, van Zwol R (2008) Flickr tag recommendation based on collective knowledge. In: WWW 327–336
31. Sivic J, Zisserman A (2003) Video google: a text retrieval approach to object matching in videos. In: ICCV 2:1470–1477
32. Smeaton AF, Over P, Kraaij W (2006) Evaluation campaigns and TRECVid. In: MIR 321–330
33. Sun Y, Kojima A (2011) A novel method for semantic video concept learning using web images. In: MM 1081–1084
34. Ulges A, Koch M, Borth D (2012) Linking visual concept detection with viewer demographics. In: ICMR. doi:10.1145/2324796.2324827
35. Ulges A, Schulze C, Keysers D, Breuel T (2008) A system that learns to tag videos by watching youtube. In: ICVS 5008:415–424
36. Ulges A, Schulze C, Keysers D, Breuel T (2008) Identifying relevant frames in weakly labeled videos for training concept detectors. In: CIVR 9–16
37. Uricchio T, Ballan L, Bertini M, Del Bimbo A (2013) An evaluation of nearest-neighbor methods for tag refinement. In: ICME 1–6
38. van de Sande K, Gevers T, Snoek CGM (2010) Evaluating color descriptors for object and scene recognition. In: PAMI 32(9):1582–1596
39. Vapnik VN (1998) Statistical learning theory. Wiley, New York
40. Wang H, Schmid C (2013) Action recognition with improved trajectories. In: ICCV 3551–3558
41. Wang Z, Zhao M, Song Y, Kumar S, Li B (2010) Youtubecat: learning to categorize wild web videos. In: CVPR
42. Yan R, Hauptmann AG, Jin R (2003) Negative pseudo-relevance feedback in content-based video retrieval. I:n MM 343–346
43. Yang J, Hauptmann A (2008) (Un)reliability of video concept detection. In: CIVR 85–94
44. Zhao W-L, Wu X, Ngo C-W (2010) On the annotation of web videos by efficient near-duplicate search. In: TMM 12(5):448–461
45. Zhu S, Ngo C-W, Jiang Y-G (2012) Sampling and ontologically pooling web images for visual concept learning. In: TMM 14(4):1068–1078

Svetlana Kordumova received Ir. diploma in computer science from the University "Ss. Cyril and Methodius", Skopje, Macedonia in 2010, where she has been awarded twice as one of the best students. For her graduation thesis she worked on gender identification and attention sensing at Philips Laboratories in Eindhoven, The Netherlands. Svetlana started a PhD in mining online multimedia at the University of Amsterdam, The Netherlands in 2011. She is part of the award-winning MediaMill Semantic Video Search Engine, which is a consistent top performer in the yearly NIST TRECVID evaluations.



Xirong Li received the B.Sc. and M.Sc. degrees from the Tsinghua University, China, in 2005 and 2007, respectively, and the Ph.D. degree from the University of Amsterdam, The Netherlands, in 2012, all in computer science. The title of his thesis is “Content-based visual search learned from social media”. He is currently an assistant professor at the Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China. His research interest is image search and multimedia content analysis. He is recipient of the SIGMM Best PhD Thesis Award 2013, the IEEE Transactions on Multimedia Prize Paper Award 2012, Best Paper Nominee of the ACM International Conference on Multimedia Retrieval 2012, Chinese Government Award for Outstanding Self-Financed Students Abroad 2011, and the Best Paper Award of the ACM International Conference on Image and Video Retrieval 2010. He served as publicity co-chair for ICMR 2013.

Cees G. M. Snoek received the M.Sc. degree in business information systems (2000) and the Ph.D. degree in computer science (2005), both from the University of Amsterdam, Amsterdam, The Netherlands. He is currently an Associate Professor in the Intelligent Systems Lab at the University of Amsterdam.

In addition, he is head of R&D at Euvision Technologies, one of the lab’s spin-off. He was previously at Carnegie Mellon University, USA (2003) and UC Berkeley’s (2010–2011). His research interests focus on video and image retrieval.

Dr. Snoek is the lead researcher of the award-winning MediaMill Semantic Video Search Engine, which is a consistent top performer in the yearly NIST TRECVID evaluations. He has published over 100 refereed book chapters, journal and conference papers. He is co-initiator and co-organizer of the annual VideOlympics, co-chair of: ACM Multimedia 2016, SPIE Multimedia Content Access conference 2010–2013, and program co-chair of the International Workshop on Content-Based Multimedia Indexing 2010, the ACM International Conference on Internet Multimedia Computing and Services 2013, and the Pacific-Rim Conference on Multimedia 2014. He is a senior member of ACM and IEEE. Dr. Snoek is member of the editorial boards for IEEE MultiMedia and IEEE Transactions on Multimedia. Cees is recipient of an NWO Veni award (2008), a Fulbright Junior Scholarship (2010), an NWO Vidi award (2012), and the Netherlands Prize for ICT Research (2012). Several of his Ph.D. students have won best paper awards, including the IEEE Transactions on Multimedia Prize Paper Award (2012) and the SIGMM Award for Outstanding PhD Thesis in Multimedia Computing, Communications and Applications (2013).